Painting by evolutionary history: Inference of local ancestry in admixed populations

Ali J. Berens (ali.berens@biology.gatech.edu), Joseph Lachance
School of Biology, Georgia Institute of Technology

Background

Local ancestry inference is important for:

- Understanding history of populations
- Personalized medicine

STRUCTURE-like approaches do not paint local ancestry

Most existing methods require reference sequences

Objective: ***Develop a reference-free, hierarchical local ancestry inference method***

Test Data

1000 Genomes Project: ASW, CEU, YRI populations
Phased chromosomes
SNP filtering: Biallelic and MAF > 0.05

Workflow

For full dataset:

Chromosomal Ancestry Estimate (ADMIXTURE, k=2)

→ Select Informative SNPs

→ Create haplotypes of l consecutive SNPs

→ Create emission matrix for each window of length l

Q-file

	A1	A2
Sample 1 (C1)	0.763	0.247
Sample 2 (C2)	0.155	0.845
⋮	⋮	⋮

Phased Chromosomes

	C1	…	Cj
SNP 1	0	⋯	1
⋮	⋮	⋱	⋮
SNP i	1	⋯	0

ALT sums weighted by **Q-file** → Top frequency differences

	A1	A2
SNP 1	0.10	0.15
SNP 2	0.33	0.24
⋮	⋮	⋮

Haplotypes (l =5)

	C1	…	Cj
SNPs 1−5	00101	⋯	01100
SNPs 10 − 15	⋮	⋱	⋮
⋮	10001	⋯	00001

SNPs 1-5

C1	00101
⋮	⋮
Cj	01100

haplotype sums weighted by **Q-file** → Emission matrix

	00000	…	11111
A1	0.674	…	0.002
A2	0.622	…	0.001

For each sample:

Create transition matrix

→ Forward HMM, Backward HMM

→ Consensus Local Ancestry Estimate

Transition matrix

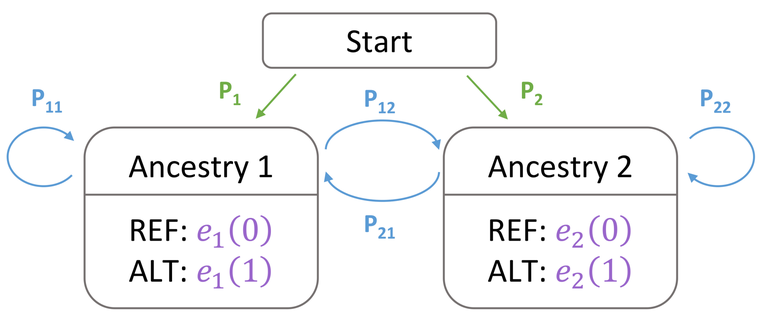
Previous State \ Next State	A1	A2
A1	$1 - rp_2$	rp_2
A2	rp_1	$1 - rp_1$

r: recombination rate
p_1, p_2: from **Q-file**

Hidden Markov Model

Example: For l = 1

Hidden: Sequence of local ancestry states, $S = \{s_1, s_2, \dots, s_i, \dots\}$
Observe: Sequence of haplotypes, $\mathrm{X} = \{x_1, x_2, \dots, x_i, \dots\}$



Hidden Markov Model (cont.)

Initialize (Forward Direction):

$\mathrm{P}_1(x_1, 1) = \log(\mathrm{P}_1) + \log(e_1(x_1))$ $\quad \mathrm{P}_2(x_1, 1) = \log(\mathrm{P}_2) + \log(e_2(x_1))$

Iterate:

$$\mathrm{P}_1(x_i, i) = \log(e_1(x_i)) + \max[\log(\mathrm{P}_{11}) + \log(\mathrm{P}_1(x_{i-1}, i-1)), \log(\mathrm{P}_{21}) + \log(\mathrm{P}_2(x_{i-1}, i-1))]$$

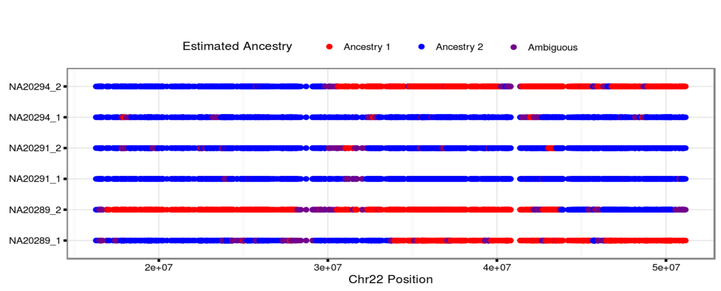
$$\mathrm{P}_2(x_i, i) = \log(e_2(x_i)) + \max[\log(\mathrm{P}_{12}) + \log(\mathrm{P}_1(x_{i-1}, i-1)), \log(\mathrm{P}_{22}) + \log(\mathrm{P}_2(x_{i-1}, i-1))]$$

State selection:

If $\mathrm{P}_1(x_i, i) > \mathrm{P}_2(x_i, i)$, then $S_i = \text{Ancestry 1}$
If $\mathrm{P}_2(x_i, i) > \mathrm{P}_1(x_i, i)$, then $S_i = \text{Ancestry 2}$

Results

Figure 1: Local ancestry estimates for homologous chromosomes from 3 ASW samples. Each sample has some discrepancy between forward and backward estimates. Yet, summed local ancestry estimate close to Admixture results.



Summed Local Ancestry	
1	2
0.610	0.390
0.026	0.974
0.026	0.974
0.002	0.998
0.774	0.226
0.610	0.390

Admixture (k=2) Ancestry	
1	2
0.661	0.339
0.001	0.999
0.033	0.966
0.001	0.999
0.738	0.262
0.628	0.372

Figure 2: Summation of local ancestry estimates highly correlated with Admixture (k=2) results

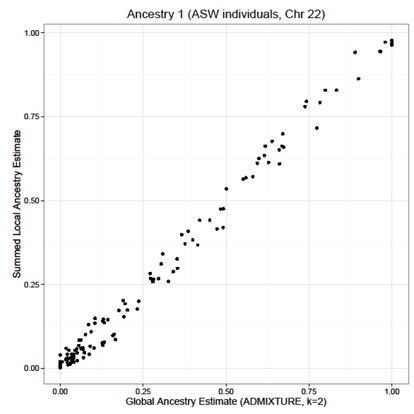
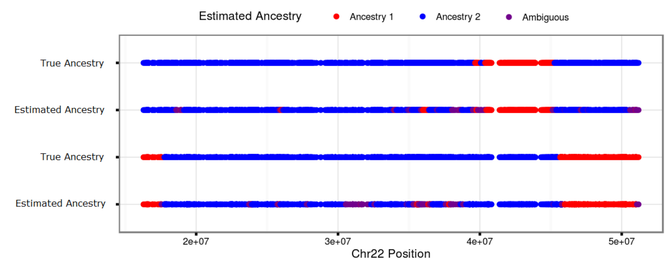


Figure 3: Local ancestry estimates for known mixture between CEU and YRI ancestry.



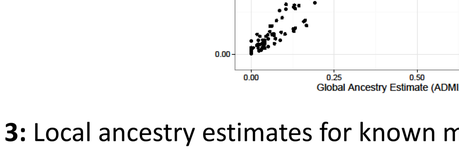
Summed Local Ancestry	
1	2
0.239	0.761
0.254	0.746
0.147	0.853
0.158	0.842

Future Improvements

1. Reduce short, erroneous ancestral flips
   a. Sliding windows of haplotypes

Selection based on:
1. Number of windows with ancestry
2. Weighted by size of haplotype
3. Weighted by confidence in ancestry call

   b. Incorporate recombination maps
   c. Post-processing of ambiguous local ancestry estimates



Depending on:
1. Length of ambiguous region
2. Relative confidence in ancestry call by direction

2. Inference with greater than 2 ancestral populations
3. Hierarchical painting of ancestry

Inference methods assume no evolutionary relationship

Evolutionary relationships do exist between populations